Optimal Data Acquisition for Statistical Estimation

YILING CHEN, Harvard University, USA
NICOLE IMMORLICA, Microsoft Research New England, USA
BRENDAN LUCIER, Microsoft Research New England, USA
VASILIS SYRGKANIS, Microsoft Research New England, USA
JUBA ZIANI, California Institute of Technology, USA

We consider a data analyst's problem of purchasing data from strategic agents to compute an unbiased estimate of a statistic of interest. Agents incur private costs to reveal their data and the costs can be *arbitrarily correlated* with their data. Once revealed, data are verifiable. This paper focuses on *linear* unbiased estimators. We design an individually rational and incentive compatible mechanism that optimizes the worst-case mean-squared error of the estimation, where the worst-case is over the unknown correlation between costs and data, subject to a budget constraint in expectation. We characterize the form of the optimal mechanism in closed-form. We further extend our results to acquiring data for estimating a parameter in regression analysis, where private costs can correlate with the values of the dependent variable but not with the values of the independent variables.

Full Paper: https://arxiv.org/abs/1711.01295

CCS Concepts: • **Theory of computation → Algorithmic mechanism design**;

Additional Key Words and Phrases: buying data; budget-feasible mechanism design; statistical estimation; optimization

1 INTRODUCTION

In the age of automation, data is king. The statistics and machine learning algorithms that help curate our online content, diagnose our diseases, and drive our cars, among other things, are all fueled by data. Typically, this data is mined by happenstance: as we click around on the internet, seek medical treatment, or drive "smart" vehicles, we leave a trail of data. This data is recorded and used to make estimates and train machine learning algorithms. So long as representative data is readily abundant, this approach may be sufficient. But some data is sensitive and therefore inaccurate, rare, or lacking detail in observable data traces. In such cases, it is more expedient to buy the necessary data directly from the population.

Consider, for example, the problem a public health administration faces in trying to learn the average weight of a population, perhaps as an input to estimating the risk of heart disease. Weight

Yiling Chen was partially supported by NSF grant CCF-1718549. Juba Ziani was supported by NSF grants CNS-1331343 and CNS-1518941, and the US-Israel Binational Science Foundation grant 2012348. Part of the work was done while Yiling Chen and Juba Ziani were at Microsoft Research New England.
Authors' addresses: Yiling Chen, Harvard University, Cambridge, MA, USA, yiling@seas.harvard.edu; Nicole Immorlica, Microsoft Research New England, Cambridge, MA, USA, nicimm@microsoft.com; Brendan Lucier, Microsoft Research New England, Cambridge, MA, USA, brlucier@microsoft.com; Vasilis Syrgkanis, Microsoft Research New England, Cambridge, MA, USA, vasy@microsoft.com; Juba Ziani, California Institute of Technology, Pasadena, CA, USA, jziani@caltech.edu.

ACM EC'18, June 18–22, 2018, Ithaca, NY, USA. ACM ISBN 978-1-4503-4529-3/18/06…$15.00
https://doi.org/10.1145/3219166.3219195

is a sensitive personal characteristic, and people may be loath to disclose it. It is also variable over time, and so must be collected close to the time of the average weight estimate in order to be accurate. Thus, while other characteristics, like height, age, and gender, are fairly accurately recorded in, for example, driver's license databases, weight is not. The public health administration may try surveying the public to get estimates of the average weight, but these surveys are likely to have low response rates and be biased towards healthier low-weight samples.

In this paper, we propose a mechanism for buying verifiable data from a population in order to estimate a statistic of interest, such as the expected value of some function of the underlying data. We assume each individual has a private cost, or disutility, for revealing his or her sensitive data to the analyst. Importantly, this cost may be correlated with the private data. For example, overweight or underweight individuals to have a higher cost of revealing their data than people of a healthy weight. Individuals wish to maximize their expected utility, which is the expected payment they receive for their data minus their expected cost. The analyst has a fixed budget for buying data. The analyst does not know the distribution of the data: properties of the distribution is what she is trying to learn from the data samples, therefore it is important that she uses the data she collects to learn it rather than using an inaccurate prior distribution (for example, the analyst may have a prior on weight distribution within a population from DMV records or previous surveys, but such a prior may be erroneous if people do not accurately report their weights). However, we do assume the analyst has a prior for the marginal distribution of costs, and that she estimates how much a survey may cost her as a function of said prior.[1]

The analyst would like to buy data subject to her budget, then use that data to obtain an unbiased estimator for the statistic of interest. To this end, the analyst posts a menu of probability-price pairs. Each individual i with cost c selects a pair (A, P) from the menu, at which point the analyst buys the data with probability A at price P. The expected utility of the individual is thus $(P - c)A$.[2] To form an estimate based on this collected data, we assume the analyst uses *inverse propensity scoring*, pioneered by Horvitz and Thompson [13]. This is the unique unbiased linear estimator; it works by upweighting the data from individual i by the inverse of his/her selection probability, $1/A$.

The Horvitz-Thompson estimator always generates an unbiased estimate of the statistic being measured, regardless of the price menu. However, the precision of the estimator, as measured by the variance or mean-squared error of the estimate, depends on the menu of probability-price pairs offered to each individual. For example, offering a high price would generate data samples with low bias (since many individuals would accept such an offer), but the budget would limit the number of samples. Offering low prices allows the mechanism to collect more samples, but these would be more heavily biased, requiring more aggressive correction which introduces additional noise. The goal of the analyst is to strike a balance between these forces and post a menu that minimizes the variance of her estimate in the worst-case over all possible joint distributions of the data and cost consistent with the cost prior. We note that this problem setting was first studied by Roth and Schonebeck [21], who characterized an approximately optimal mechanism for moment estimation.

1.1 Summary of results and techniques

Our main contribution comes in the form of an exact solution for the optimal menu, as discussed in Section 3. As one would expect, if the budget is large, the optimal menu offers to buy, with probability 1, all data at a cost equal to the maximum cost in the population. If the budget is

[1] This prior could come from similar past exercises. Alternatively, when no prior is known, the analyst can allocate a fraction of his budget to buying data for the sake of learning this distribution of costs. In this paper, we follow prior work (e.g., Roth and Schoenebeck [21]) and assume that a prior distribution is known, instead of focusing on how one might learn the distribution of costs.

[2] As we show, this menu-based formulation is fully general and captures arbitrary data-collection mechanisms.

small, the optimal menu buys data from an individual with probability inversely proportional to the square root of their cost.[3] Interestingly, in intermediate regimes, we show the optimal menu employs pooling: for all individuals with sufficiently low private cost, it buys their data with equal probability; for the remaining high cost agents, it buys their data with probability inversely proportional to the square root of their costs. Revisiting the example of estimating the weight of a population of size n, our scheme suggests the following solution. Imagine the costs are $0, 4, 8$ with probability $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$, and the total budget of the analyst is $B = 7n$. The analyst brings a scale to a public location and posts the following menu of pairs of allocation probability and price: $\{(1, \frac{36}{5}), (\frac{4}{5}, 8)\}$. A simple calculation shows that individuals with cost 0 or 4 will pick the first menu option: stepping on the scale and having their weight recorded with probability 1, and receiving a payment of $\frac{36}{5}$ dollars. Individuals with cost 8 will pick the second menu option; if they are selected to step on the scale, which happens with probability $\frac{4}{5}$, the analyst records their weight scaled by a factor of $\frac{5}{4}$. This scaling is precisely the upweighting from inverse propensity scoring. In expectation over the population, the analyst spends exactly his budget $7n$. The estimate is the average of the scaled weights.

We show how to extend our approach in multiple directions. First, our characterization of the optimal mechanism holds even when the quantity to be estimated is the expected value of a d-dimensional moment function of the data. Second, we extend our techniques beyond moment estimation to the common task of multi-dimensional linear regression. In this regression problem, an individual's data includes both features (which are assumed to be insensitive or publicly available) and outcomes (which may be sensitive). The analyst's goal is to estimate the linear regression coefficients that relate the outcomes to the features. We make the assumption that an individual's cost is independent of her features, but may be arbitrarily correlated with her outcome. For example, the goal might be to regress a health outcome (such as severity of a disease) on demographic information. In this case, we might imagine that an agent incurs no cost for reporting his age, height or gender, but his cost might be highly correlated with his realized health outcome. In such a setting, we show that the asymptotically optimal allocation rule, given a fixed average budget per agent as the number of agent grows large, can be calculated efficiently and exhibits a pooling region as before. However, unlike for moment estimation, agents with intermediate costs can also be pooled together. We further show that our results extend to non-linear regression in the full version of the paper, under mild additional conditions on the regression function.

Our techniques rely on i) reducing the mechanism design problem to an optimization problem through the classical notion of virtual costs, then ii) reducing the problem of optimizing the worst-case variance to that of finding an equilibrium of a zero-sum game between the analyst and an adversary. The adversary's goal is to pick a distribution of data, conditional on agents' costs, that maximizes the variance of the analyst's estimator. We then characterize such an equilibrium through the optimality conditions for convex optimization described in [2].

1.2 Related work

A growing amount of attention has been placed on understanding interactions between the strategic nature of data holders and the statistical inference and learning tasks that use data collected from these holders. The work on this topic can be roughly divided into two categories according to whether money is used for incentive alignment.

[3] Of course, the individual is him/herself selecting the menu option and so the use of an active verb in this context is perhaps a bit misleading. What we mean here is that, given his/her incentives based on his/her private cost, the choice the individual selects is one that buys his/her data with probability inversely proportional to the square root of his/her cost.

In the first category, individuals as data holders do not directly derive utility from the accuracy of the inference or learning outcome, but in some cases may incur a privacy cost if the outcome leaks their private information. The analyst uses monetary payments to incentivize agents to reveal their data. Our work falls into this category. Prior papers by Roth and Schoenebeck [21] and Abernethy et al. [1] are closest to our setting. Similarly to our work, both Roth and Schoenebeck [21] and Abernethy et al. [1] consider an analyst's problem of purchasing data from individuals with private costs subject to a budget constraint, allow the cost to be correlated with the value of data, and assume that individuals cannot fabricate their data. Roth and Schoenebeck [21] aim at obtaining an optimal unbiased estimator with minimum worst-case variance for population mean, while their mechanism achieves optimality only approximately: instead of the actual worst-case variance, a bound on the worst-case variance is minimized. While our setting is identical to that of [21], our work precisely minimizes worst-case variance (under a regularity assumption on the cost distribution), and our main contribution is to exhibit the structure of the optimal mechanism, as well as to extend our results to broader classes of statistical inference, moment estimation and linear regression. In particular, compared to [21], our solution exhibits new structure in the form of a pooling region for low cost agents; i.e., the optimal mechanism pools agents with the lowest costs together and treats them identically. Such structure does not arise in [21] under a regularity assumption on the cost distribution. Abernethy at al. [1] consider general supervised learning. They do not seek to achieve a notion of optimality; instead, they take a learning-theoretic approach and design mechanisms to obtain learning guarantees (risk bounds).

Several papers consider data acquisition models with different objectives under the assumptions that (a) individuals do not fabricate their data, and (b) private costs and value of data are uncorrelated. For example, in the work of Cummings et al. [6], the analyst can decide the level of accuracy for data purchased from each individual, and wishes to guarantee a certain desired level of accuracy of the aggregated information while minimizing the total privacy cost incurred by the agents. Cai et al. [3] focus on incentivizing individuals to exert effort to obtain high-quality data for the purpose of linear regression. Another line of research in the first category examines the data acquisition problem under the lens of differential privacy [5, 9–11, 18]. The mechanism designer then uses payments to balance the trade-off between privacy and accuracy.

In the second category, individuals' utilities directly depend on the inference or learning outcome (e.g. they want a regression line to be as close to their own data point as possible) and hence they have incentives to manipulate their reported data to influence the outcome. There often is no cost for reporting one's data. The data analyst, without using monetary payments, attempts to design or identify inference or learning processes so that they are robust to potential data manipulations. Most papers in this category assume that independent variables (feature vectors) are unmanipulable public information and dependent variables are manipulable private information [7, 14, 15, 19], though some papers consider strategic manipulation of feature vectors [8, 12]. Such strategic data manipulations have been studied for estimation [4], classification [12, 14, 15], online classification [8], regression [7, 20], and clustering [19]. Work in this category is closer to mechanism design without money in the sense that they focus on incentive alignment in acquiring data (e.g., strategy-proof algorithms) but often do not evaluate the performance of the inference or learning, with a few notable exceptions [8, 12].

2 MODEL AND PRELIMINARIES

Survey Mechanisms. There is a population of n agents. Each agent i has a private pair (z_i, c_i), where $z_i \in \mathcal{Z}$ is a data point and $c_i > 0$ is a cost. We think of c_i as the disutility agent i incurs by releasing her data z_i. The pair is drawn from a distribution $\mathcal{D}$, unknown to the mechanism designer.

We denote with $\mathcal{F}$ the CDF of the marginal distribution of costs,[4] supported on a set C. We assume that $\mathcal{F}$ and the support of the data points, $\mathcal{Z}$, are known. However, the joint distribution $\mathcal{D}$ of data and costs is unknown.

A *survey mechanism* is defined by an allocation rule $A : C \to [0, 1]$ and a payment rule $P : C \to \mathbb{R}$, and works as follows. Each agent i arrives at the mechanism in sequence and reports a cost $\hat{c}_i$. The mechanism chooses to buy the agent's data with probability $A(\hat{c}_i)$. If the mechanism buys the data, then it learns the value of z_i (i.e., agents cannot misreport their data) and pays the agent $P(\hat{c}_i)$. Otherwise the data point is not learned and no payment is made.

We assume agents have quasi-linear utilities, so that the utility enjoyed by agent i when reporting $\hat{c}_i$ is

$$u(\hat{c}_i; c_i) = (P(\hat{c}_i) - c_i) \cdot A(\hat{c}_i) \tag{1}$$

We will restrict attention to survey mechanisms that are truthful and individually rational.

Definition 1 (Truthful and Individually Rational - TIR). *A survey mechanism is truthful if for any cost c it is in the agent's best interest to report their true cost, i.e. for any report $\hat{c}$:*

$$u(c; c) \geq u(\hat{c}; c) \tag{2}$$

It is individually rational if, e. for any cost $c \in C$, $P(c) \geq c$.

We assume that the mechanism is constrained in the amount of payment it can make to the agents. We will formally define this as an expected budget constraint for the survey mechanism.

Definition 2 (Expected Budget Constraint). *A mechanism respects a budget constraint B if:*

$$n \cdot \mathbb{E}_{c \sim \mathcal{F}}[P(c) \cdot A(c)] \leq B \tag{3}$$

Estimators. The designer (or *data analyst*) wishes to use the survey mechanism to estimate some parameter $\theta \in \mathbb{R}$ of the marginal distribution of data points.[5] For example, it might be that $\mathcal{Z} = \mathbb{R}$ and θ is the mean of the distribution over data points in the population. To this end, the designer will apply an *estimator* to the collection of data points S elicited by the survey mechanism. We will write $\hat{\theta}_S$ for the estimator used. Note that the value of the estimator $\hat{\theta}_S$ depends on the sample S, but might also depend on the distribution of costs $\mathcal{F}$ and the survey mechanism. Due to the randomness inherent in the survey mechanism (both in the choice of data points sampled and the values of those samples), we think of $\hat{\theta}_S$ as a random variable, drawn from a distribution $\mathcal{T}(\mathcal{D}, A)$. We will focus exclusively on *unbiased* estimators.

Definition 3 (Unbiased Estimator). *Given an allocation function A, an estimator $\hat{\theta}_S$ for θ is* unbiased *if for any instantiation of the true distribution $\mathcal{D}$ its expected value is equal to θ:*

$$\mathbb{E}_{\hat{\theta}_S \sim \mathcal{T}(\mathcal{D}, A)}\left[\hat{\theta}_S\right] = \theta. \tag{4}$$

Given a fixed choice of estimator, the mechanism designer wants to construct the survey mechanism to minimize the variance (finite sample or asymptotic as the population grows) of that estimator. Since the designer does not know the distribution $\mathcal{D}$, we will work with the worst-case variance over all instantiations of $\mathcal{D}$ that are consistent with the cost marginal $\mathcal{F}$.

Definition 4 (Worst-Case Variance). *Given an allocation function A and an instance of the true distribution $\mathcal{D}$, the* variance *of an estimator $\hat{\theta}_S$ is defined as:*

$$\mathbb{V}(\hat{\theta}_S; \mathcal{D}, A) = \mathbb{E}_{\hat{\theta}_S \sim \mathcal{T}(\mathcal{D}, A)}\left[\left(\hat{\theta}_S - \mathbb{E}\left[\hat{\theta}_S\right]\right)^2\right] \tag{5}$$

[4] Throughout the text we will use the CDF to refer to the distribution itself.

[5] We also extend our results to multi-dimensional parameters; see Section 4.

The worst-case variance *of* $\hat{\theta}_S$ *is*

$$\mathbb{V}^*(\hat{\theta}_S; \mathcal{F}, A) = \sup_{\mathcal{D} \text{ consistent with } \mathcal{F}} \mathbb{V}(\hat{\theta}_S; \mathcal{D}, A). \tag{6}$$

We are now ready to formally define the mechanism design problem faced by the data analyst.

Definition 5 (Analyst's Mechanism Design Problem). *Given an estimator* $\hat{\theta}_S$ *and cost distribution* $\mathcal{F}$*, the goal of the designer is to design an allocation rule* A *and payment rule* P *so as to minimize worst-case variance subject to the truthfulness, individual rationality and budget constraints:*

$$\begin{aligned} &\inf_{A,P} \ \mathbb{V}^*(\hat{\theta}_S; \mathcal{F}, A) \\ &\text{s.t. } n \cdot \mathbb{E}_{c\sim\mathcal{F}}[P(c) \cdot A(c)] \le B \\ &\qquad A, P \text{ define a TIR mechanism} \end{aligned} \tag{7}$$

Implementing Surveys as Posted Menus. The formulation above describes surveys as direct-revelation mechanisms, where agents report costs. We note that an equivalent indirect implementation might be more natural: a *posted menu survey* offers each agent a menu of (price, probability) pairs $(p_1, A_1), \ldots, (p_k, A_k)$. If the agent chooses (p_m, A_m) then their data is elicited with probability A_m, in which case they are paid p_m. Each agent can choose the item that maximizes their expected utility, i.e., $\text{argmax}_{m\in[k]} (p_m - c) \cdot A_m$. By the well-known *taxation principle*, any survey mechanism can be implemented as a posted menu survey, and the number of menu items required is at most the size of the support of the cost distribution.

2.1 Reducing Mechanism Design to Optimization

We begin by reducing the mechanism design problem to a simpler full-information optimization problem where the designer knows the private cost of each player and can acquire their data by paying them exactly that cost. However, the designer is constrained to using *monotone* allocation rules, in which players with higher costs have weakly lower probability of being chosen.

Definition 6 (Analyst's Optimization Problem). *Given an estimator* $\hat{\theta}_S$ *and cost distribution* $\mathcal{F}$*, the optimization version of the designer's problem is to find a non-increasing allocation rule* A *that minimizes worst-case variance subject to the budget constraint, assuming agents are paid their cost:*

$$\begin{aligned} &\inf_{A} \ \mathbb{V}^*(\hat{\theta}_S; \mathcal{F}, A) \\ &\text{s.t. } n \cdot \mathbb{E}_{c\sim\mathcal{F}}[c \cdot A(c)] \le B \\ &\qquad A \text{ is monotone non-increasing} \end{aligned} \tag{8}$$

The mechanism design problem in Definition 5 reduces to the optimization problem given by Definition 6, albeit with a transformation of costs to *virtual cost.*

Definition 7 (Virtual Costs and Regular Distributions). *If* $\mathcal{F}$ *is continuous and admits a density* f *then define the virtual cost function as* $\phi(c) = c + \frac{\mathcal{F}(c)}{f(c)}$*. If* $\mathcal{F}$ *is discrete with support* $C = \{c_1, \ldots, c_{|C|}\}$ *and PDF* f*, then define the virtual cost function as:* $\phi(c_t) = c_t + \frac{c_t - c_{t-1}}{f(c_t)}\mathcal{F}(c_{t-1})$*, with* $c_0 = 0$*. We also denote with* $\phi(\mathcal{F})$ *the distribution of virtual costs; i.e., the distribution created by first drawing* c *from* $\mathcal{F}$ *and then mapping it to* $\phi(c)$*. A distribution* $\mathcal{F}$ *is regular if the virtual cost function is increasing.*

When $\mathcal{F}$ is twice-continuously differentiable, $\mathcal{F}$ is regular if and only if $\mathcal{F}(c)f'(c) < 2f(c)^2$ for all $c \in C$. Importantly, in this case, the allocation rule of Roth and Schoenebeck [21] is monotone strictly decreasing in c and does not exhibit a pooling region at low-cost as our solution does.

The following is an analogue of Myerson's [16] reduction of mechanism design to virtual welfare maximization, adapted to the survey design setting.

Lemma 2.1. *If the distribution of costs $\mathcal{F}$ is regular, then solving the Analyst's Mechanism Design Problem reduces to solving the Analyst's Optimization Problem for distribution of costs $\phi(\mathcal{F})$.*

Proof. The proof is given in the full version of the paper. □

2.2 Unbiased Estimation and Inverse Propensity Scoring

We now describe a class of estimators $\hat{\theta}_S$ that we will focus on for the remainder of the paper. Note that simply calculating the quantity of interest, θ, on the sampled data points can lead to bias, due to the potential correlation between costs and data. For instance, suppose that $z \in \mathbb{R}$ and the goal is to estimate the mean of the distribution of z. A natural estimator is the average of the collected data: $\hat{\theta}_S = \frac{1}{|S|}\sum_{i\in S} z_i$. However, if players with lower z tend to have lower cost, and are therefore selected with higher probability by the analyst, then this estimator will consistently underestimate the true mean.

This problem can be addressed using *inverse propensity scoring* (IPS), pioneered by Horvitz and Thompson [13]. The idea is to recover unbiasedness by weighting each data point by the inverse of the probability of observing it. This IPS approach can be applied to any parameter estimation problem where the parameter of interest is the expected value of an arbitrary *moment function* $m : \mathcal{Z} \to \mathbb{R}$.

Definition 8 (Horvitz-Thompson Estimator). *The Horvitz-Thompson estimator for the case when the parameter of interest is the expected value of a (moment) function $m : \mathcal{Z} \to \mathbb{R}$ is defined as:*

$$\hat{\theta}_S = \frac{1}{n}\sum_{i\in[n]} \frac{m(z_i)\cdot 1\{i \in S\}}{A(c_i)} \tag{9}$$

The Horvitz-Thompson estimator is the unique unbiased estimator that is a linear function of the observations $m(z_i)$ [21]. It is therefore without loss of generality to focus on this estimator if one restricts to unbiased linear estimators.[6]

IPS beyond moment estimation. We defined the Horvitz-Thompson estimator with respect to moment estimation problems, $\theta = \mathbb{E}[m(z)]$. As it turns out, this approach to unbiased estimation extends even beyond the moment estimation problem to parameter estimation problems defined as the solution to a system of moment equations $\mathbb{E}[m(z;\theta)] = 0$ or parameters defined as the minima of a moment function $\operatorname{argmin}_\theta \mathbb{E}[m(z;\theta)]$. We defer this discussion to Section 5.

3 ESTIMATING MOMENTS OF THE DATA DISTRIBUTION

In this section we consider the case where the analyst's goal is to estimate the mean of a given moment function of the distribution. That is, there is some function $m : C \to [0,1]$ such that both 0 and 1 are in the support of random variable $m(z)$, and the goal of the analyst is to estimate $\theta = \mathbb{E}[m(z)]$.[7] We assume that $\hat{\theta}_S$, the estimator being applied, is the Horvitz-Thompson estimator given in Definition 8.

For convenience we will assume that the cost distribution $\mathcal{F}$ has finite support, say $C = \{c_1, \ldots, c_{|C|}\}$ with $c_1 < \ldots < c_{|C|}$. (We relax the finite support assumption in Section 3.2.) Write

[6] We note that we have assumed, for convenience, that $A(c_i) > 0$ for all $i \in [n]$ in the expression of this estimator, for it to be unbiased and well-defined. It is easy to see from the expression for the variance given in Section 3 that the variance-minimizing allocation rule will indeed be non-zero for each cost.

[7] Observe that it is easy to deal with the more general case of $m(z) \in [a, b]$ by a simple linear translation, i.e., estimate $\tilde{m}(z) = \frac{m(z)-a}{b-a}$ instead, which is in [0, 1] and then translate the estimator back to recover $m(z)$.

$\pi_t = f(c_t)$ for the probability of cost c_t in $\mathcal{F}$. Also, for a given allocation rule A, we will write $A_t = A(c_t)$ for convenience. That is, we can interpret an allocation rule A as a vector of $|C|$ values $A_1, \ldots, A_{|C|}$. For further convenience, we will write $q_t = \Pr[m(z) = 1|c_t]$. This is the probability that the moment takes on its maximum value when the cost is c_t. Finally, we will assume that the distribution of costs is regular.

Our goal is to address the analyst's mechanism design problem for this restricted setting. By Lemma 2.1 it suffices to solve the analyst's optimization problem. We start by characterizing the worst-case variance for this setting.

LEMMA 3.1. *The worst-case variance of the Horvitz-Thompson estimator of a moment $m : C \to [0, 1]$, given cost distribution $\mathcal{F}$ and allocation rule A, is:*

$$n \cdot \mathbb{V}^*(\hat{\theta}_S; \mathcal{F}, A) = \sup_{q \in [0,1]^{|C|}} \sum_{t=1}^{|C|} \pi_t \cdot \frac{q_t}{A_t} - \left(\sum_{t=1}^{|C|} \pi_t \cdot q_t \right)^2 \tag{10}$$

PROOF. For any distribution $\mathcal{D}$, observe that the Horvitz-Thompson estimator can be written as the sum of n i.i.d. random variables each with a variance:

$$\sigma^2 = \mathbb{E}\left[\left(\frac{m(z_i) \cdot 1\{i \in S\}}{A(c_i)} \right)^2 \right] - \mathbb{E}\left[\frac{m(z_i) \cdot 1\{i \in S\}}{A(c_i)} \right]^2 = \sum_{t=1}^{|C|} \pi_t \cdot \frac{\mathbb{E}\left[m(z)^2 | c_t\right]}{A_t} - \mathbb{E}[m(z)]^2$$

Hence, the variance of the estimator is σ^2/n. Observe that conditional on any value c, the worst-case distribution $\mathcal{D}$, will assign positive mass only to values $z \in \mathcal{Z}$ such that $m(z) \in \{0, 1\}$. This is because any other conditional distribution can be altered by a mean-preserving spread, pushing all the mass on these values, while preserving the conditional mean $\mathbb{E}[m(z)|c]$. This would strictly increase the latter variance. Thus we can assume without loss of generality that $m(z) \in \{0, 1\}$, in which case $m(z)^2 = m(z)$ and $\mathbb{E}[m(z)|c] = \Pr[m(z) = 1|c]$. Recall that $q_t = \Pr[m(z) = 1|c_t]$. Then we can simplify the variance as:

$$n \cdot \mathbb{V}(\hat{\theta}_S; \mathcal{D}, A) = \sum_{t=1}^{|C|} \pi_t \cdot \frac{\mathbb{E}[m(z)|c_t]}{A_t} - \mathbb{E}[m(z)]^2 = \sum_{t=1}^{|C|} \pi_t \cdot \frac{q_t}{A_t} - \left(\sum_{t=1}^{|C|} \pi_t \cdot q_t \right)^2$$

The theorem follows since the worst-case variance is a supremum over all possible consistent distributions, hence equivalently a supremum over conditional probabilities $q : [0, 1]^{|C|}$. □

Given the above characterization of the variance of the estimator, we can greatly simplify the analyst's optimization problem for this setting. Indeed, it suffices to find the allocation rule $A \in (0, 1]^{|C|}$ that minimizes (10), subject to A being monotone non-decreasing and satisfying the expected budget constraint.

3.1 Characterization of the Optimal Allocation Rule

We are now ready to solve the analyst's optimization problem for moment estimation. In this and all following sections, we denote $\bar{B} = \frac{B}{n}$ for simplicity of notations, and refer to $\bar{B}$ as the "average budget per agent". Note that different agents with different costs may be allocated different fractions of the total budget B that in general do not coincide with $\bar{B}$. We remark that if $\bar{B}$ is larger than the expected cost of an agent, then it is feasible (and hence optimal) for the analyst to set the allocation rule to pick any type with probability 1. We therefore assume without loss of generality that $\mathbb{E}[c] > \bar{B}$.

Our analysis is based on an equilibrium characterization, where we view the analyst choosing A and the adversary choosing z as playing a zero-sum game and solve for its equilibria. We first

present the characterization and some qualitative implications and then present an outline of our proof. We defer the full details of the proof to the full version of the paper.

Theorem 3.2 (Optimal Allocation for Moment Estimation). *The optimal allocation rule A is determined by two constants $\bar{A}$ and $t^* \in \{0, \ldots, |C|\}$ such that:*

$$A_t = \begin{cases} \bar{A} & \text{if } t \le t^* \\ \frac{\alpha}{\sqrt{c_t}} & \text{o.w.} \end{cases} \tag{11}$$

with α uniquely determined such that the budget constraint is binding.[8] Moreover, the parameters $\bar{A}$ and t^* can be computed in time $O(\log(|C|))$.

The parameters $\bar{A}$ and t^* in Theorem 3.2 are explicitly derived in closed form in the full version of the paper. For instance, when $\bar{B} \ge \frac{c_{|C|}}{2}$, then $t^* = |C|$ and $A_t = \bar{A} = \min\left\{1, \bar{B}/\mathbb{E}[c]\right\}$ for all t. When $\bar{B} \le \frac{\sqrt{c_1}\,\mathbb{E}[\sqrt{c}]}{2}$ then $t^* = 0$ and $A_t = \frac{\bar{B}}{\sqrt{c_t}\,\mathbb{E}[\sqrt{c}]}$. In fact, it can be shown (see full version) that in this latter case, the worst-case distribution is given by $q = 1$. In particular, in this restricted case, the approximation of Roth and Schoenebeck [21] is in fact optimal, and indeed our allocation rule is expressing the solution of Roth and Schoenebeck [21] as a posted menu for a discrete, regular distribution of costs. In every other case, $q \neq 1$ and our solution differs from that of Roth and Schoenebeck [21], exhibiting a pooling region for low-cost agents. More generally, the computational part of Theorem 3.2 follows by performing binary search over the support of $\mathcal{F}$, which can be done in $O(\log(|C|))$ time.

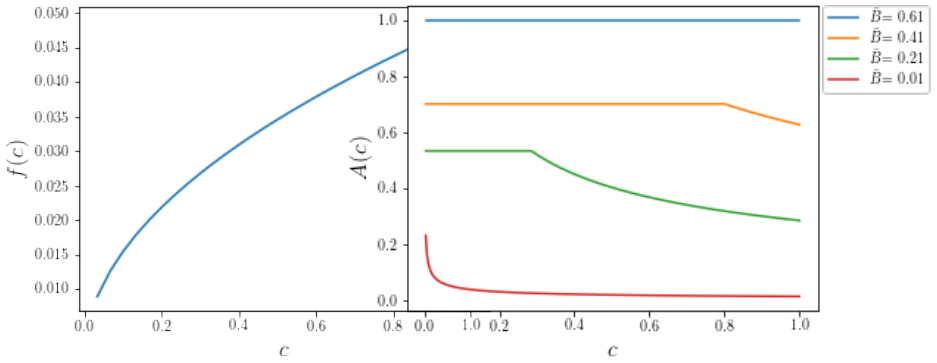

Fig. 1. The pdf (left) of a distribution of costs and the corresponding optimal allocation rule for varying levels of per-agent budget (right). Note that for sufficiently large budgets, a flat pooling region forms for agents with low costs.

We note that the optimal rule essentially allocates to each agent inversely proportionally to the square root of their cost, but may also "pool" the allocation probability for agents at the lower end of the cost distribution. See Figure 1 for examples of optimal solutions.

The proof of Theorem 3.2 appears in the full version of the paper. The main idea is to view the optimization problem as a zero-sum game between the analyst who designs the allocation rule A, and an adversary who designs q so as to maximize the variance of the estimate. We show how to compute an equilibrium (A^*, z^*) of this zero-sum game via Lagrangian and KKT conditions, and then note that the obtained A^* must in fact be an optimal allocation rule for worst-case variance.

[8] The explicit form of this is $\alpha = \frac{\bar{B} - \bar{A}\,\mathbb{E}[c \cdot 1\{c \le c_{t^*}\}]}{\mathbb{E}[\sqrt{c} \cdot 1\{c > c_{t^*}\}}$.

The analysis above applied to a discrete cost distribution over a finite support of possible costs. We show how to extend this analysis to a continuous distribution below, noting that the continuous variant of the Optimization Problem for Moment Estimation can be derived by taking the limit over finer and finer discrete approximations of the cost distribution.

3.2 Continuous Costs for Moment Estimation

Definition 9 (Continuous Optimization Problem for Moment Estimation). *When costs are supported on $C = [0, 1]$, the analyst's optimization problem for the moment estimation problem based on the Horvitz-Thompson estimator can be written as:*

$$\inf_{A:(0,1]\to[0,1]} \sup_{x:[0,1]\to[0,1]} \int_0^1 \frac{x(c)}{A(c)}\, d\mathcal{F}(c) - \left(\int_0^1 x(c)\, d\mathcal{F}(c)\right)^2$$
$$\text{s.t.} \int_0^1 c \cdot A(c)\, d\mathcal{F}(c) \le \bar{B} \tag{12}$$
$$A \text{ is monotone non-increasing}$$

We can now establish the following continuous variant of Theorem 3.2, which describes the optimal survey mechanism for continuous cost distributions.

Theorem 3.3 (Continuous Limit of Optimal Allocation). *If the distribution of costs is atomless and supported in $(0, 1]$, then the optimal allocation rule A is determined by two constants $\bar{A}$ and $x^* \in \mathbb{R}$ such that:*

$$A(c) = \begin{cases} \bar{A} & \text{if } c \le x^* \\ \frac{\alpha}{\sqrt{c}} & \text{o.w.} \end{cases} \tag{13}$$

with α uniquely determined such that the budget constraint is binding.[9]*. The quantities $\bar{A}$ and x^* are defined as follows: for any $x \in \mathbb{R}$ let*

$$Q_\infty(x) = \mathbb{E}_{c\sim\mathcal{F}}[\min\{c, \sqrt{cx}\}] \qquad R_\infty(x) = 2\,\mathbb{E}_{c\sim\mathcal{F}}\left[\min\left\{\frac{c}{x}, 1\right\}\right] \qquad G(x) = \frac{Q_\infty(x)}{\max(1, R_\infty(x))}$$

Then $x^* = \min\{1, G^{-1}(\bar{B})\}$ and $\bar{A} = \frac{1}{\max(1, R_\infty(x^*))}$ (see Figure 2).[10]

Proof. See full version. □

Let us give some intuition behind the form of the allocation rule described in Theorem 3.3. As in Theorem 3.2, the allocation rule will pool agents with low costs (i.e., less than some threshold x^*), then allocate to higher-cost agents inversely proportional to the square root of their costs. In the definition of x^* and $\bar{A}$, note that Q_∞ is non-decreasing and R_∞ is non-increasing, so G is non-decreasing. We therefore have that x^*, the boundary of the pooling region, increases with $\bar{B}$ up to a maximum value of 1 (at which point all agents are pooled).

Let's restrict attention to the case where the mean of the distribution is at least as large as half of the maximum value of the support, i.e. $\mathbb{E}[c] \ge 1/2$. In this setting, we see that $R_\infty(x) \le 1$ for all $x \in [0, 1]$, so

$$G(x) = \frac{Q_\infty(x)}{R_\infty(x)} = \frac{x}{2} \cdot \frac{\mathbb{E}_{c\sim\mathcal{F}}[\min\{c, \sqrt{cx}\}]}{\mathbb{E}_{c\sim\mathcal{F}}[\min\{c, x\}]} \approx \frac{x}{2} \qquad \text{(see Figure 2)}$$

So the optimal allocation sets $x^* \approx 2\bar{B}$. Moreover, the allocation for the pooling region is $\bar{A} = \frac{1}{R_\infty(x^*)} \approx \frac{\bar{B}}{\mathbb{E}[\min\{c, 2\bar{B}\}]}$. So the optimal mechanism takes the following intuitive form: first, assign

[9] The explicit form of this is $\alpha = \frac{\bar{B} - \bar{A}\,\mathbb{E}[c\cdot 1\{c\le x*\}]}{\mathbb{E}[\sqrt{c}\cdot 1\{c > x^*\}]}$.

[10] We take the convention that if $\bar{B}$ lies above the range of G, then $G^{-1}(\bar{B}) = +\infty$.

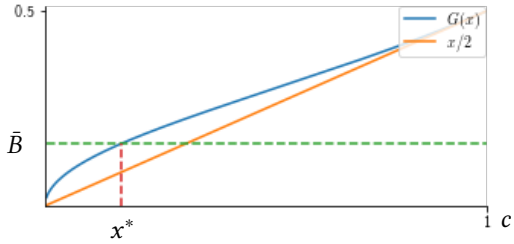

Fig. 2. Pictorial representation of the function $G(x)$ which defines the optimal threshold x^*, portraying its close approximation by the function $x/2$.

each agent an allocation probability $\bar{A}$ that would, in an alternate world where costs are capped at $2\bar{B}$, precisely exhaust the budget. Since costs can actually be greater than $2\bar{B}$, this flat allocation goes over-budget. So, for agents whose costs are greater than $2\bar{B}$, we remove allocation probability so that (a) the budget becomes balanced, and (b) the remaining probability of allocation is inversely proportional to the square root of the costs.

4 MULTI-DIMENSIONAL PARAMETERS FOR MOMENT ESTIMATION

Section 3 focused on the case of estimating a single-dimensional parameter of the data distribution. In this section we note that our characterization of the optimal mechanism extends to multi-dimensional moment estimation as well. In multi-dimensional moment estimation, there is a function $m\colon C \to [0,1]^d$, and our goal is to estimate $\theta = \mathbb{E}[m(z)] \in [0,1]^d$. Here d is the dimension of the estimation problem, which we assume to be a fixed constant.

As before, we will estimate θ by applying an estimator $\hat{\theta}$ to the data collected from a survey mechanism. To evaluate an estimator, we must extend our definition of variance to the d-dimensional setting, as follows.

Definition 10 (Worst-Case Mean Squared Error - Risk). *Given allocation function A and distribution $\mathcal{D}$, the expected mean squared error (or* risk*) of an estimator $\hat{\theta}$ is*

$$\mathcal{R}(\hat{\theta}_S; \mathcal{D}, A) = \mathbb{E}_{\hat{\theta}_S \sim \mathcal{T}(\mathcal{D},A)} \left[\left\| \hat{\theta}_S - \theta_0 \right\|_2^2 \right] \tag{14}$$

and the worst-case variance of $\hat{\theta}$ is

$$\mathcal{R}^*(\hat{\theta}_S; \mathcal{F}, A) = \sup_{\mathcal{D} \text{ consistent with } \mathcal{F}} \mathcal{R}(\hat{\theta}_S; \mathcal{D}, A). \tag{15}$$

When $\hat{\theta}$ is unbiased, the risk has a natural interpretation: it is simply the sum of variances of each coordinate of θ, considered separately.

Claim 1 (Risk of Unbiased Estimators). *The risk of any unbiased estimator is equal to the sum of variances of every coordinate:*

$$\mathcal{R}(\hat{\theta}_S; \mathcal{D}, A) = \mathbb{E}_{\hat{\theta}_S \sim \mathcal{T}(\mathcal{D},A)} \left[\sum_{r=1}^{d} (\hat{\theta}_{S,r} - \mathbb{E}[\hat{\theta}_{S,r}])^2 \right] = \sum_{r=1}^{d} \mathbb{V}(\hat{\theta}_{S,r}). \tag{16}$$

As in the single-dimensional case, the analyst obtains an estimate through the Horvitz-Thompson estimator, which is defined as follows for parameters in $\mathbb{R}^d$. Also as in the single-dimensional case, The Horvitz-Thompson estimator is an unbiased estimator of $\mathbb{E}[m(z)]$.

Definition 11 (Horvitz-Thompson Estimator for Multi-dimensional Moment Estimation). *The Horvitz-Thompson estimator for the case when the parameter of interest is the expected value of a vector of moments $m : \mathcal{Z} \to \mathbb{R}^d$ is defined as:*

$$\hat{\theta}_S = \frac{1}{n} \sum_{i \in [n]} \frac{1\{i \in S\}}{A(c_i)} \cdot m(z_i) \tag{17}$$

For our characterization of worst-case risk, we will assume that the moment function m can take on the extreme points of the hypercube $[0,1]^d$.

Assumption 1. *$\mathcal{D}$ is such that the induced distribution of $m(z)$ is supported on every extreme point of the $[0,1]^d$ hypercube.*

Lemma 4.1. *Under Assumption 1, the worst-case risk of the Horvitz-Thompson estimator of moment $m : C \to [0,1]$ is*

$$\frac{n}{d} \cdot \mathcal{R}^*(\hat{\theta}_S; \mathcal{F}, A) = \sup_{q \in [0,1]^{|C|}} \sum_{t=1}^{|C|} \pi_t \cdot \frac{q_t}{A_t} - \left(\sum_{t=1}^{|C|} \pi_t \cdot q_t \right)^2 . \tag{18}$$

Proof. See full version. □

Lemma 4.1 implies that the optimal survey design problem in the d-dimensional case is, in fact, identical to the problem considered in the single-dimensional case. We can conclude that Theorems 3.2 and 3.3, which characterized the optimal survey mechanisms for discrete and continuous single-parameter settings, respectively, also apply to the multi-dimensional setting without change.

5 MULTI-DIMENSIONAL PARAMETER ESTIMATION VIA LINEAR REGRESSION

In this section, we extend beyond moment estimation to a multi-dimensional linear regression task (we discuss the non-linear case in the full version). For this setting we will impose additional structure on the data held by each agent. Each agent's private information consists of a feature vector $x_i \in \mathbb{R}$, an outcome value $y_i \in \mathbb{R}$, and a residual value $\epsilon_i \in \mathbb{R}$, that are i.i.d among agents. Each agent also has a cost c_i. The data is generated in the following way: first, x_i is drawn from an unknown distribution $\mathcal{X}$. Then, independently from x_i, the pair (c_i, ϵ_i) is drawn from a joint distribution $\mathcal{D}$ over $\mathbb{R}^2$. The marginal distribution over costs, $\mathcal{D}_c$, is known to the designer, but not the full joint distribution $\mathcal{D}$. Then y_i is defined to be

$$y_i = x_i^\top \theta^* + \epsilon_i \tag{19}$$

where $\theta^* \in \Theta$ with Θ a compact subset of $\mathbb{R}^d$. We further require that θ^* is in the interior of Θ. We write $\mathcal{D}_\epsilon$ for the marginal distribution over ϵ_i, which is supported on some bounded range $[L, U]$, and has expected value 0. (So, in particular, $L \leq 0 \leq U$.)

When a survey mechanism buys data from agent i, the pair (x_i, y_i) is revealed. Crucially, the value of ϵ_i is not revealed to the survey mechanism. The goal of the designer is to estimate the parameter vector θ^*.

Note that the single-dimensional moment estimation problem from Section 3 is a special case of linear regression. Indeed, consider setting $d = 1$, $\epsilon_i = m(z_i) - \mathbb{E}[m(z_i)]$ for each i, $\theta^* = \mathbb{E}[m(z_i)]$, and x_i to be the constant -1. Then, when the survey mechanism purchases data from agent i, it learns $y_i = m(z_i)$, and estimating θ^* is equivalent to estimating the expected value of $m(z_i)$.

More generally, one can interpret x_i as a vector of publicly-verifiable information about agent i, which might influence a (possibly sensitive) outcome y_i. For example, x_i might consist of demographic information, and y_i might indicate the severity of a medical condition. The coefficient vector θ^* describes the average effect of each feature on the outcome, over the entire population.

Under this interpretation, ϵ_i is the residual agent-specific component of the outcome, beyond what can be accounted for by the agent's features. We can interpret the independence of x_i from (c_i, ϵ_i) as meaning that each agent's cost to reveal information is potentially correlated with their (private) residual data, but is independent of the agent's features.

As in Section 3, the analyst wants to design a survey mechanism to buy from the agents, obtain data from the set S of elicited agents, then compute an estimate $\hat{\theta}_S$ of θ. The expected average payment to each of the n agents should be no more than $\bar{B}$. As in Section 2.1, we note that the problem of designing a survey mechanism in fact reduces to that of designing an allocation rule A that minimizes said variance and satisfies a budget constraint in which the prices are replaced by known virtual costs. To this end, the analyst designs an allocation rule A and a pricing rule P so as to minimize the $\sqrt{n}$-normalized worst-case asymptotic mean-squared error of $\hat{\theta}_S$ as the population size goes to infinity. Our mechanism will essentially be optimizing the coefficient in front of the leading $1/n$ term in the mean squared error, ignoring potential finite sample deviations that decay at a faster rate than $1/n$. Note that we will design allocation and pricing rules to be independent of the population size n; hence, the analyst can use the designed mechanism even if the exact population size in unknown.

5.1 Estimators for Regression

Let S be the set of data points elicited by a survey mechanism. The analyst's estimate will then be the value $\hat{\theta}_S$ that minimizes the Horvitz-Thompson mean-squared error $\mathbb{E}[(y_i - x_i^\top \theta^*)^2]$, i.e.,

$$\hat{\theta}_S = \operatorname*{argmin}_{\theta \in \Theta} \sum_i \frac{1\{i \in S\}}{A(c_i)} (y_i - x_i^\top \theta)^2. \tag{20}$$

Further, we make the following assumptions on the distribution of data points:

Assumption 2 (Assumption on the distribution of features). *$E[x_i x_i^\top]$ is finite and positive-definite, and hence invertible.*

Finite expectation is a property one may expect real data such has age, height, weight, etc. to exhibit. The second part of the assumption is satisfied by common classes of distributions, such as multivariate normals. We first show that $\hat{\theta}_S$ is a consistent estimator of θ.

Lemma 5.1. *Under Assumption 2, for any allocation rule $A > 0$ that does not depend on n, $\hat{\theta}_S$ is a consistent estimator of θ^*.*

Proof of Lemma 5.1. Let $m(\theta; x, y) = (y - x^\top \theta)^2$, and let $w_i = 1\{i \in S\}$ for simplicity. The following holds:

(1) First, we note that θ^* is the unique parameter that minimizes $\mathbb{E}[(y_i - x_i^\top \theta)^2]$; indeed, take any $\theta \neq \theta^*$, we have that

$$\begin{aligned}\mathbb{E}[(y_i - \theta^\top x_i)^2] &= \mathbb{E}\left[\left(y_i - x_i^\top \theta^* + x_i^\top(\theta^* - \theta)\right)^2\right] \\ &= \mathbb{E}\left[\left(y_i - x_i^\top \theta^*\right)^2\right] + \mathbb{E}\left[\left(x_i^\top(\theta^* - \theta)\right)^2\right] + 2\,\mathbb{E}\left[\epsilon_i (\theta^* - \theta)^\top x_i\right]\end{aligned}$$

As x and ε are independent, ε has mean 0, this simplifies to

$$\begin{aligned}\mathbb{E}[(y_i - \theta^\top x_i)^2] &= \mathbb{E}\left[\left(y_i - \theta^{*\top} x_i\right)^2\right] + (\theta^* - \theta)^\top \mathbb{E}\left[x_i x_i^\top\right](\theta^* - \theta) + 2(\theta^* - \theta)^\top \mathbb{E}\left[\epsilon_i x_i\right] \\ &= \mathbb{E}\left[\left(y_i - \theta^{*\top} x_i\right)^2\right] + (\theta^* - \theta)^\top \mathbb{E}\left[x_i x_i^\top\right](\theta^* - \theta) \\ &> \mathbb{E}\left[\left(y_i - \theta^{*\top} x_i\right)^2\right]\end{aligned}$$

where the last step follows from $\mathbb{E}[x_i x_i^\top]$ being positive-definite by Assumption 2.

(2) By definition, Θ is compact.

(3) $m(\theta; x, y)$ is continuous in θ, and so is its expectation.

(4) $m(.;.)$ is also bounded (lower-bounded by 0, and upper-bounded by either L^2 or U^2), implying that $\theta \to \frac{w_i}{A(c_i)} m(\theta; x_i, y_i)$ is continuous and bounded. Hence, by the uniform law of large number, remembering that $\frac{w_i}{A(c_i)} m(\theta; x_i, y_i)$ are i.i.d,

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \frac{w_i}{A(c_i)} m(\theta; x_i, y_i) - \mathbb{E}\left[\frac{w_i}{A(c_i)} m(\theta; x_i, y_i) \right] \right| \to 0.$$

Finally, noting that conditional on c_i, $m(\theta; x_i, y_i)$ and $\frac{w_i}{A(c_i)}$ are independent, we have:

$$\mathbb{E}\left[\frac{w_i}{A(c_i)} m(\theta; x_i, y_i) \right] = \mathbb{E}\left[\mathbb{E}\left[\frac{w_i}{A(c_i)} \,\middle|\, c_i \right] \mathbb{E}[\, m(\theta; x_i, y_i) \mid c_i \,] \right] = \mathbb{E}\left[m(\theta; x_i, y_i) \right]$$

using $\mathbb{E}\left[\frac{w_i}{A(c_i)} \,\middle|\, c_i \right] = 1$.

Therefore, all of the conditions of Theorem 2.1 of [17] are satisfied, which is enough to prove the result. □

Similarly to the moment estimation problem in Section 3, the goal of the analyst is to minimize the worst-case (over the distribution of data and the correlation between c_i's and ϵ_i's) asymptotic mean-squared error of the estimator $\hat{\theta}_S$. Here "asymptotic" means the worst-case error as $\hat{\theta}_S$ approaches the true parameter θ^*. The following theorem characterizes the asymptotic covariance matrix of $\hat{\theta}_S$. (In fact, it fully characterizes the asymptotic distribution of $\hat{\theta}_S$.)

Lemma 5.2. *Under Assumption 2, for any allocation rule $A > 0$ that does not depend on n, the asymptotic distribution of $\hat{\theta}_S$ is given by*

$$\sqrt{n}(\hat{\theta}_S - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}[x_i x_i^\top]^{-1} \mathbb{E}\left[\epsilon_i^2 \frac{1\{i \in S\}}{A^2(c_i)} \right] \right)$$

where d denotes convergence in distribution and where randomness in the expectations is taken on the costs c_i, the set of elicited data points S, the features of the data x_i, and the noise ϵ_i.

Proof of Lemma 5.2. For simplicity, let $w_i = 1\{i \in S\}$ and note that the w_i's are i.i.d. Let $m(\theta; x_i, y_i) = (y_i - x_i^\top \theta)^2$. First we remark that $\nabla_\theta m(\theta; x_i, y_i) \cdot \frac{w_i}{A(c_i)} = 2\frac{w_i}{A(c_i)} x_i (x_i^\top \theta - y_i)$ and $\nabla^2_{\theta\theta} m(\theta; x_i, y_i) \cdot \frac{w_i}{A(c_i)} = 2\frac{w_i}{A(c_i)} x_i x_i^\top$. We then note the following:

(1) θ^* is in the interior of Θ.

(2) $\theta \to m(\theta; x_i, y_i) \cdot \frac{w_i}{A(c_i)}$ is twice continuously differentiable for all x_i, y_i, c_i, w_i.

(3) $\sqrt{n}\left(\frac{1}{n} \sum_{i=1}^n \nabla_\theta\, m(\theta^*; x_i, y_i) \cdot \frac{w_i}{A(c_i)} \right) \to \mathcal{N}\left(0, 4\,\mathbb{E}\left[\frac{w_i}{A^2(c_i)} x_i x_i^\top (x_i^\top \theta^* - y_i)^2 \right]\right)$. This follows directly from applying the multivariate central limit theorem, noting that

$$\mathbb{E}\left[\nabla_\theta m(\theta^*; x_i, y_i) \cdot \frac{w_i}{A(c_i)} \right] = \mathbb{E}\left[\mathbb{E}[\, 2x_i \epsilon_i \mid c_i \,] \cdot \mathbb{E}\left[\frac{w_i}{A(c_i)} \,\middle|\, c_i \right] \right] = \mathbb{E}\left[2x_i \epsilon_i \right] = 0$$

where the first step follows from conditional indepence on c of x, ϵ with $A(c), S$, the second step from $\mathbb{E}\left[\frac{w_i}{A(c_i)} \,\middle|\, c_i \right] = 1$, and the last equality follows from the fact that x and ϵ are independent and $\mathbb{E}[\epsilon_i] = 0$.

(4) $\sup_{\theta\in\Theta}\left\|\mathbb{E}\left[\nabla^2_{\theta\theta}m(\theta;x_i,y_i)\cdot\frac{w_i}{A(c_i)}\right]-\frac{1}{n}\sum_i\nabla^2_{\theta\theta}m(\theta;x_i,y_i)\cdot\frac{w_i}{A(c_i)}\right\|\to 0$, applying the uniform law of large numbers as $\nabla^2_{\theta\theta}m(\theta;x_i,y_i)\cdot\frac{w_i}{A(c_i)}=2\frac{w_i}{A(c_i)}x_ix_i^\top$ is i) continuous in θ, and ii) constant in θ, thus bounded coordinate-by-coordinate by $2\frac{w_i}{A(c_i)}x_ix_i^\top$ that is independent of θ and has finite expectation $2\,\mathbb{E}[x_ix_i^\top]$.

(5) $\mathbb{E}\left[\nabla^2_{\theta\theta}m(\theta;x_i,y_i)\cdot\frac{w_i}{A(c_i)}\right]=2\,\mathbb{E}[x_ix_i^\top]$ is invertible as it is positive-definite.

Therefore the sufficient conditions i)-v) in Theorem 3.1 of [17] hold, proving that the asymptotic distribution is normal with mean 0 and variance

$$\mathbb{E}[2x_ix_i^\top]^{-1}\,\mathbb{E}\left[4\frac{w_i}{A^2(c_i)}(x_i^\top\theta-y_i)^2x_ix_i^\top\right]\mathbb{E}[2x_ix_i^\top]^{-1}.$$

To conclude the proof, we remark that by independence of x_i with c_i and ε_i,

$$\mathbb{E}\left[\frac{w_i}{A^2(c_i)}(x_i^\top\theta-y_i)^2x_ix_i^\top\right]=\mathbb{E}[x_ix_i^\top]\,\mathbb{E}\left[\epsilon_i^2\frac{w_i}{A^2(c_i)}\right]$$

□

Lemma 5.2 implies that the worst-case asymptotic mean-squared error, under a budget constraint, is given by the worst-case trace of the variance matrix. That is,

$$\mathcal{R}^*(\mathcal{F},A)\triangleq\sup_{\mathcal{X}}\sup_{\mathcal{D}_\epsilon}\mathbb{E}\left[\epsilon_i^2\frac{1\{i\in S\}}{A^2(c_i)}\right]\cdot\sum_{j=1}^d\mathbb{E}[x_ix_i^\top]^{-1}_{jj}\quad\text{s.t. }\mathbb{E}[\epsilon_i]=0 \tag{21}$$

where recall that $\mathcal{D}_\epsilon$ is the marginal distribution over ϵ and $\mathcal{X}$ the distribution over x. Importantly, this can be rewritten as

$$\mathcal{R}^*(\mathcal{F},A)\triangleq\left(\sup_{\mathcal{X}}\sum_{j=1}^d\mathbb{E}[x_ix_i^\top]^{-1}_{jj}\right)\cdot\sup_{\mathcal{D}_\epsilon}\mathbb{E}\left[\epsilon_i^2\frac{1\{i\in S\}}{A^2(c_i)}\right]\quad\text{s.t. }\mathbb{E}[\epsilon_i]=0 \tag{22}$$

Therefore, the analyst's decision solely depend on the worst-case correlation between costs c_i and noise ϵ_i, and not on the worst-case distribution $\mathcal{X}$. In turn, the analyst's allocation is completely independent of and robust in $\mathcal{X}$.

5.2 Characterizing the Optimal Allocation Rule for Regression

As in Section 3, we assume costs are drawn from a discrete set, say $C=\{c_1,\dots,c_{|C|}\}$. We will then write A_t for an allocation rule conditional on the cost being c_t, and π_t the probability of the cost of an agent being c_t. We will assume that $\bar{B}<\sum_{t=1}^{|C|}\pi_tc_t$, meaning that it is not feasible to accept all data points, since otherwise it is trivially optimal to set $A_t=1$ for all t.

The following lemma describes the optimization problem faced by an analyst wanting to design an optimal survey mechanism. Recall that residual values lie in the interval $[L,H]$.

Lemma 5.3 (Optimization Problem for Parameter Estimation). *The optimization program for the analyst is given by:*

$$\begin{aligned} \inf_{A\in[0,1]^{|C|}} \sup_{q\in[0,1]^{|C|}} & \sum_{t=1}^{l} \frac{\pi_t}{A_t}\left((1-q_t)\cdot L^2 + q_t\cdot U^2\right) \\ \text{s.t.} & \sum_{t=1}^{|C|} \pi_t\left((1-q_t)\cdot L + q_t\cdot U\right) = 0 \\ & \sum_{t=1}^{|C|} \pi_t c_t A_t \le \bar{B} \\ & A \text{ is monotone non-increasing} \end{aligned} \tag{23}$$

Proof of Lemma 5.3. First we note that

$$\begin{aligned} \mathcal{R}^*(\mathcal{F}, A) \triangleq \left(\sup_{X} \sum_{j=1}^{d} \mathbb{E}[x_i x_i^\top]^{-1}_{jj}\right) \cdot \sup_{\mathcal{D}_\epsilon} & \mathbb{E}\left[\epsilon_i^2 \frac{1\{i\in S\}}{A^2(c_i)}\right] \\ \text{s.t. } & \mathbb{E}[\epsilon_i] = 0 \end{aligned} \tag{24}$$

We can therefore renormalize the worst-case variance by $\sup_X \sum_{i=1}^{d} \mathbb{E}[x_i x_i^\top]^{-1}_{ii}$, as it does not depend on any other parameter of the problem. The analyst's objective is now given by

$$\begin{aligned} \sup_{\mathcal{D}_\epsilon} & \sum_{t=1}^{|C|} \pi_t \frac{\mathbb{E}[\epsilon_i^2|c_t]}{A_t} \\ \text{s.t. } & \mathbb{E}[\epsilon_i] = 0 \end{aligned} \tag{25}$$

The worst case distribution is reached when $\epsilon_i|c_t$ is binomial between L and U (and such a distribution is feasible for $\epsilon_i|c_t$), therefore letting $q_t = P[\epsilon_i = U \mid c_t]$, we obtain the lemma. □

We can now characterize the form of the optimal survey mechanism. For simplicity, we will assume that $U^2 \ge L^2$. This is without loss of generality, since the optimization program is symmetric in L and U; if $L^2 > U^2$, the analyst can set $q_t = 1 - q_t$, $L = U$ and $U = L$ to obtain Program (23) with $U^2 > L^2$.

Theorem 5.4. *Under the assumptions described above, an optimal allocation rule A has the form*

(1) $A_t = \min\left(1, \alpha\frac{|L|}{\sqrt{c_t}}\right)$ *for* $t < t^-$
(2) $A_t = \bar{A}$ *for all* $t \in \{t^-, \dots, t^+\}$
(3) $A_t = \min\left(1, \alpha\frac{U}{\sqrt{c_t}}\right)$ *for* $t > t^+$

for $\bar{A}$ and α positive constants that do not depend on n, and t^- and t^+ integers with $t^- \le t^+$. Further, $\bar{A}$ and α can be computed efficiently given knowledge of t^-, t^+.

We remark that the allocation rule that we designed is strictly positive and independent of n (as the optimization program itself does not depend on n), so Lemmas 5.1 and 5.2 apply. Theorem 5.4 immediately implies that an optimal allocation rule can be obtained by simply searching over the space of parameters (t^-, t^+), which can be done in at most $|C|^2$ steps. For each pairs of parameters (t^-, t^+), A can be computed efficiently as stated in the Theorem. Then the analyst only needs to pick the allocation rule that minimizes the objective value among the obtained allocation rules that are feasible for Program (23). Further, we remark that the solution for the linear regression case exhibits a structure that is similar to the structure of the optimal allocation rule for moment

estimation (see Theorem 3.2): it exhibits a pooling region in which all cost types are treated the same way, and changes in the inverse of the square root of the cost outside said pooling region. However, we note that we may now choose to pool agents together in an intermediate range of costs, instead of pooling together agents whose costs are below a given threshold.

Proof sketch. We first compute the best response q^* of the adversary; we note that this best response is in fact the solution to a knapsack problem that is independent of the value taken by the allocation rule A. We can therefore plug the adversary's best response into the optimization problem, and reduce the minimax problem above in a simple minimization problem on A. We then characterize the solution as a function of the parameters $(t^-, t^+) \in [|C|]^2$ through KKT conditions. The full proof is given in the full version of this paper. □

Non-linear regression: We further show in the full version of the paper that our results extend to non-linear regression, i.e. when y_i is generated by a process of the more general form

$$y_i = f(\theta^*, x_i) + \epsilon_i,$$

under a few additional assumptions on the distribution of x and on the regression function f.

REFERENCES

[1] Jacob Abernethy, Yiling Chen, Chien-Ju Ho, and Bo Waggoner. 2015. Low-Cost Learning via Active Data Procurement. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation (EC '15)*. ACM, New York, NY, USA, 619–636. https://doi.org/10.1145/2764468.2764519

[2] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press, New York, NY, USA.

[3] Y. Cai, C. Daskalakis, and C. H. Papadimitriou. 2015. Optimum Statistical Estimation with Strategic Data Sources. In *Proceedings of the 28th Conference on Learning Theory*. 280–296.

[4] Ioannis Caragiannis, Ariel D. Procaccia, and Nisarg Shah. 2016. Truthful Univariate Estimators. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16)*. JMLR.org, 127–135. http://dl.acm.org/citation.cfm?id=3045390.3045405

[5] Rachel Cummings, Stratis Ioannidis, and Katrina Ligett. 2015. Truthful Linear Regression. In *Proceedings of the 28th Conference on Learning Theory*. 448–483.

[6] Rachel Cummings, Katrina Ligett, Aaron Roth, Zhiwei Steven Wu, and Juba Ziani. 2015. Accuracy for Sale: Aggregating Data with a Variance Constraint. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science (ITCS '15)*. ACM, New York, NY, USA, 317–324. https://doi.org/10.1145/2688073.2688106

[7] O. Dekel, F. Fischer, and A. D. Procaccia. 2010. Incentive Compatible Regression Learning. *J. Comput. System Sci.* 76, 8 (2010), 759–777.

[8] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. 2017. Strategic Classification from Revealed Preferences. (2017).

[9] Lisa Fleischer and Yu-Han Lyu. 2012. Approximately Optimal Auctions for Selling Privacy when Costs are Correlated with Data. *CoRR* abs/1204.4031 (2012). http://arxiv.org/abs/1204.4031

[10] Arpita Ghosh, Katrina Ligett, Aaron Roth, and Grant Schoenebeck. 2014. Buying Private Data without Verification. *CoRR* abs/1404.6003 (2014). http://arxiv.org/abs/1404.6003

[11] Arpita Ghosh and Aaron Roth. 2011. Selling Privacy at Auction. In *Proceedings of the 12th ACM Conference on Electronic Commerce (EC '11)*. ACM, New York, NY, USA, 199–208. https://doi.org/10.1145/1993574.1993605

[12] M. Hardt, N. Megiddo, C. H. Papadimitriou, and M. Wootters. 2016. Strategic Classification. In *7th*. 111–122.

[13] D. G. Horvitz and D. J. Thompson. 1952. A Generalization of Sampling Without Replacement from a Finite Universe. *J. Amer. Statist. Assoc.* 47, 260 (1952), 663–685. https://doi.org/10.1080/01621459.1952.10483446 arXiv:http://amstat.tandfonline.com/doi/pdf/10.1080/01621459.1952.10483446

[14] R. Meir, S. Almagor, A. Michaely, and J. S. Rosenschein. 2011. Tight bounds for strategyproof classification. In *10th*. 319–326.

[15] R. Meir, A. D. Procaccia, and J. S. Rosenschein. 2012. Algorithms for Strategyproof Classification. *Artificial Intelligence* 186 (2012), 123–156.

[16] Roger B. Myerson. 1981. Optimal Auction Design. *Math. Oper. Res.* 6, 1 (Feb. 1981), 58–73. https://doi.org/10.1287/moor.6.1.58

[17] Whitney K. Newey and Daniel McFadden. 1986. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, R. F. Engle and D. McFadden (Eds.). Handbook of Econometrics, Vol. 4. Elsevier, Chapter 36, 2111–2245. https://ideas.repec.org/h/eee/ecochp/4-36.html

[18] Kobbi Nissim, Salil Vadhan, and David Xiao. 2014. Redrawing the Boundaries on Purchasing Data from Privacy-sensitive Individuals. In *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science (ITCS '14)*. ACM, New York, NY, USA, 411–422. https://doi.org/10.1145/2554797.2554835

[19] J. Perote and J. Perote-Peña. 2003. The impossibility of strategy-proof clustering. *Economics Bulletin* 4, 23 (2003), 1–9.

[20] J. Perote and J. Perote-Peña. 2004. Strategy-proof estimators for simple regression. *Mathematical Social Sciences* 47 (2004), 153–176.

[21] Aaron Roth and Grant Schoenebeck. 2012. Conducting Truthful Surveys, Cheaply. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC '12)*. ACM, New York, NY, USA, 826–843. https://doi.org/10.1145/2229012.2229076